Statistical End-to-end Performance Bounds for Networks under Long Memory FBM Cross Traffic

Amr Rizk and Markus Fidler

Institute of Communications Technology, Leibniz Universität Hannover

Abstract—**Fractional Brownian motion (fBm) emerged as a useful model for self-similar and long-range dependent Internet traffic. Approximate performance measures are known from large deviations theory for single queuing systems with fBm through traffic. In this paper we derive end-to-end performance bounds for a through flow in a network of tandem queues under fBm cross traffic. To this end, we prove a rigorous sample path envelope for fBm that complements previous approximate results. We find that both approaches agree in their outcome that overflow probabilities for fBm traffic have a Weibullian tail. We employ the sample path envelope and the concept of leftover service curves to model the remaining service after scheduling fBm cross traffic at a system. Using composition results for tandem systems from the stochastic network calculus we derive end-to-end statistical performance bounds for individual flows in networks under fBm cross traffic. We discover that these bounds grow in $\mathcal{O}\big(n(\log n)^{\frac{1}{2-2H}}\big)$ for n systems in series where H is the Hurst parameter of the fBm cross traffic. We show numerical results on the impact of the variability and the correlation of fBm traffic on network performance.**

I. Introduction

Since the beginning of the Internet classical queuing theory has been used as the decisive methodology to analyze many relevant problems in computer networking [1], most prominently to prove the efficiency of packet switching over circuit switching. The significance of queuing theory stems from explicit closed-form performance measures, e.g. backlog and delay, for queuing systems under Poisson traffic[1], such as the M|M|1 single server system with infinite buffer. In practice results obtained for systems with infinite buffer are frequently transferred with reasonable precision to systems with finite buffer, e.g. for buffer sizing [2]. The outstanding features of queuing theory that are essential for analysis of networks are (*i*) multiplexing and demultiplexing of flows[2], i.e. routing, and (*ii*) product form queuing networks where constituent queues can be analyzed as if in isolation.

While it has long been known that Internet traffic sources do not fulfill the memoryless property assumed by the Poisson traffic model, it has been argued that the aggregate of a large number of multiplexed flows will tend towards Poisson. This assumption is backed up by the fact that the sum of many independent Bernoulli trials converges to a Poisson random variable. In the mid 90's, the groundbreaking discovery from extensive measurements was, however, that aggregate Internet traffic possesses significant long-range dependence (LRD) as

well as statistical self-similarity [3], [4], [5], [6], [7]. A possible mathematical justification for these traffic properties arises from the superposition of many independent on-off sources with heavy-tailed on and off periods [8]. This model matches file size distributions on storage systems [5], [6].

A random process that captures self-similarity is fractional Brownian motion (fBm) [9]. FBm is characterized by the Hurst parameter H where fBm with $H \in \left(\frac{1}{2}, 1\right)$ is LRD and widely accepted in the literature [3], [10], [11] as a useful model for aggregate Internet traffic. The cases $H = \frac{1}{2}$ and $H \in \left(0, \frac{1}{2}\right)$ correspond to standard Brownian motion and short-range dependent fBm, respectively. Using theories including effective bandwidths and large deviations significant results have been derived for the performance of single queuing systems fed with LRD fBm traffic [11], [12], [13], [14]. Regarding the analysis of networks these theories can, however, not carry the outstanding properties of queuing theory forward to fBm.

Properties similar to (*i*) and (*ii*) from queuing theory have, however, been established in the deterministic [15], [16] and stochastic [15], [17], [18], [19], [20], [21] network calculus. The network calculus uses the concept of service curves [22], [23], [24] to characterize the service provided by queuing systems. Leftover service curves are used to analyze the effects of scheduling, i.e. multiplexing and demultiplexing cross traffic (*i*). Service curves of tandem systems are composed by convolution enabling the analysis of networks (*ii*). Finally, arrival envelopes are used as a model of traffic flows [25], [26], [27], [28], [29], [17] to derive performance bounds on backlog and delay for service curve systems.

In this paper we derive end-to-end statistical performance bounds for tandem systems under LRD fBm cross traffic. We contribute a rigorous sample path envelope for fBm traffic that complements an approximate envelope that follows from a known asymptotic backlog bound. Both envelopes agree in the Weibullian decay of overflow probabilities. These envelopes are the basis for network analysis using the stochastic network calculus. Owing to the concept of leftover service curves we quantify the effects of fBm cross traffic on the performance of through flows and find that the correlation of the cross traffic has severe impact. Finally, we derive end-to-end performance bounds for a through flow that traverses n tandem systems each under fBm cross traffic. We show that these bounds grow in $\mathcal{O}\big(n(\log n)^{\frac{1}{2-2H}}\big)$. Our finding compares to [19] where end-to-end performance bounds in $\mathcal{O}(n \log n)$ have been derived for traffic with exponentially as opposed to Weibull bounded burstiness. For $H = \frac{1}{2}$ we recover the previous result.

---

[1]Poisson traffic refers to exponential packet inter-arrival and service times.

[2]Multiplexing independent Poisson flows leads to a Poisson flow as well as blind demultiplexing of a Poisson flow leads to Poisson flows.

The remainder of this paper is structured as follows. In Sect. II we review related work on the application of fBm for performance analysis of computer networks. In Sect. III we provide a sample path envelope for LRD fBm traffic based on its moment generating function. In Sect. IV we derive a leftover service curve under fBm cross traffic. In Sect. V we present end-to-end service curves and statistical performance bounds for tandem systems. Sect. VI gives brief conclusions.

II. RELATED WORK ON FBM QUEUING SYSTEMS

FBm is a widely accepted model for self-similar Internet traffic [3], [10], [11]. It can be short- or long-range dependent conditional on its Hurst parameter $H \in (0, \frac{1}{2})$ or $H \in (\frac{1}{2}, 1)$, respectively. An fBm process $Z(t)$ has stationary Gaussian increments and the following basic properties: $Z(0) = 0$, $\mathsf{E}[Z(t)] = 0$, and $\mathsf{E}[Z(t)^2] = \sigma^2 t^{2H}$ for all $t \geq 0$ where $\sigma > 0$ is its standard deviation at $t = 1$. The increment process of fBm, also called fractional Gaussian noise (fGn), has positive as well as negative increments. While this property of fGn is not matched by real network traffic it avoids excessive complexity and keeps the model solvable [14]. The long-range dependence of fBm for $H \in (\frac{1}{2}, 1)$ is established by the infinite sum $\sum_t v(t) = \infty$ of the auto-covariance of the increments $v(t) \approx \sigma^2 H(2H - 1)t^{2H-2}$ as $t \to \infty$.

In the sequel cumulative arrivals from a traffic source in an interval $[\tau, t)$ are denoted $A(\tau, t)$. Shorthand notation $A(t)$ is used for $A(0, t)$. The arrivals of an fBm traffic source are modeled as the superposition of a mean rate λ and an fBm process $Z(t)$

$$A(t) = \lambda t + Z(t). \tag{1}$$

The challenge of the fBm traffic model in case of LRD is that its variance $\sigma^2 t^{2H}$ grows superlinearly in t, i.e. the traffic is heavily bursty with burst periods that are more likely to be sustained for a long time. These properties make the calculation of performance bounds for fBm traffic hard.

The backlog process $B(t)$ at a lossless work-conserving constant rate server with capacity C is described by Reich's equation, see e.g. [30],

$$B(t) = \sup_{\tau \in [0,t]} \{A(\tau, t) - C(t - \tau)\}.$$

The difficulty behind the analysis of a statistical bound b for the steady state backlog B, i.e. letting $t \to \infty$, is to find the value τ^* that achieves the supremum in

$$\mathsf{P}[B > b] = \mathsf{P}\left[\sup_{\tau \in [0,t]} \{A(\tau, t) - C(t - \tau)\} > b\right] \tag{2}$$

since τ^* is a random variable, see [17] for explanation.

Large deviations theory is frequently used to analyze the asymptotic decay rate of the overflow probability of a backlog bound [31], [12], [13], [15]. The asymptotic calculation makes use of the principle of the largest term stating

$$\mathsf{P}[B > b] \approx \sup_{\tau \in [0,t]} \mathsf{P}[A(\tau, t) - C(t - \tau) > b] \tag{3}$$

where the term on the right hand side strictly provides only a lower bound. For fBm traffic (1) at a server with capacity C the following asymptotic holds for the decay rate of the overflow probability [12]

$$\lim_{b \to \infty} \frac{\log \mathsf{P}[B > b]}{b^{2-2H}} = -\inf_{k>0} k^{2H-2} \frac{(k + C - \lambda)^2}{2}.$$

It follows that $\lim_{b \to \infty} \mathsf{P}[B > b] = \varepsilon_a$ where [12]

$$\varepsilon_a = \exp\left(-\frac{1}{2\sigma^2}\left(\frac{C-\lambda}{H}\right)^{2H}\left(\frac{b}{1-H}\right)^{2-2H}\right). \tag{4}$$

The resulting overflow probability has a Weibull tail that simplifies to an exponential distribution for the special case $H = \frac{1}{2}$. The backlog bound was proven to be logarithmically asymptotical generally and exact for $H = \frac{1}{2}$.

The large deviations result (4) agrees with a solution deduced for the largest term (3) in [10], [11]. The derivation makes use of the Gaussian distribution of the increments of fBm and yields the approximation

$$\mathsf{P}[B > b] \approx \sup_{t \geq 0} \bar{\bar{\Phi}}\left(\frac{(C-\lambda)t + b}{\sigma t^H}\right)$$

where $\bar{\bar{\Phi}}(x) = P(Z(1) > x)$ is the complementary cumulative distribution function of a Gaussian random variable, i.e. the increment of fBm. After maximizing over t the backlog bound approximation in [10] is $\mathsf{P}[B > b] \approx \varepsilon_a$ where ε_a is identical to (4). A comprehensible introduction covering the derivation of this bound can also be found in [32].

The proof of the large deviations theory builds on the Gärtner-Ellis condition which establishes a direct relation to the effective bandwidth of a traffic flow [12]. The theory of effective bandwidths, e.g. [14], [15], is a major tool for the analysis of traffic flows as it gives a measure for resource requirements at different time scales. The effective bandwidth of a flow $\alpha(\theta, t) = \frac{1}{\theta t} \log \mathsf{E}\left[e^{\theta A(t)}\right]$ lies between its average and peak rate depending on the parameter $\theta > 0$. For fBm it holds that

$$\alpha(\theta, t) = \lambda + \frac{\theta \sigma^2}{2} t^{2H-1}. \tag{5}$$

In case of $H \in (\frac{1}{2}, 1)$ the effective bandwidth of fBm traffic exhibits a continuous growth in t due to LRD [14].

In [17] a connection between effective bandwidths and effective envelopes is established. In contrast to asymptotic results for large buffers from large deviations theory, effective envelopes in conjunction with the stochastic network calculus [28], [29], [17], [18], [19], [21] can provide non-asymptotic performance bounds. Moreover, recent stochastic network calculus provides methods for derivation of stochastic leftover service curves as well as for composition of tandem systems. Effective envelopes $E(t - \tau)$ are statistical upper bounds of the cumulative arrivals $A(\tau, t)$ of the form

$$\mathsf{P}[A(\tau, t) - E(t - \tau) > 0] \leq \varepsilon_p.$$

An envelope for fBm traffic is derived in [33], [34], [17] as

$$E(t) = \lambda t + \sqrt{-2 \log \varepsilon_p} \sigma t^H. \tag{6}$$

The definition of effective envelope is point-wise in the sense that it can be violated at each point in time with overflow probability ε_p. Applying the approximation by the largest term (3) the envelope (6) is used in [34] to recover the backlog bound from large deviations theory (4).

In contrast, the derivation of performance bounds using the stochastic network calculus builds on sample path arguments, such as the backlog bound (2), and requires a bound for $A(\tau, t)$ for all $\tau \in [0, t]$, i.e. a sample path envelope of the form

$$\mathsf{P}\left[\sup_{\tau \in [0,t]} \{A(\tau, t) - E(t - \tau)\} > 0\right] \le \varepsilon_s. \quad (7)$$

Such sample path envelopes are constructed in [17] using Boole's inequality under the assumption of a time scale limit T, i.e. by summing the constant point-wise overflow probabilities ε_p over all $t \in [0, T]$. The time scale in this context can be regarded as a constraint on the duration of busy periods. In case of fBm the duration of busy periods has, however, been found to grow extremely fast with H [33].

Methods for construction of sample path envelopes that do not require a priori assumptions on the relevant time scale have been developed in [24], [29], [19]. The general approach is to use a point-wise envelope with parameter b

$$\mathsf{P}\left[A(\tau, t) - E(t - \tau) > b\right] \le \varepsilon_p(b). \quad (8)$$

that has a decaying and integrable overflow profile $\varepsilon_p(b)$, i.e. $\int_0^\infty \varepsilon_p(b)db$ is finite. Typically, when constructing a sample path envelope b is substituted by a slack rate $\varrho \cdot (t - \tau)$. The slack rate relaxes the envelope such that ε_p decreases with increasing interval width $(t - \tau)$. Finally, taking Boole's inequality over all τ to derive the sample path overflow probability ε_s (7) translates to integrating $\int_0^t \varepsilon_p(\varrho(t - \tau))d\tau$ that remains finite for all $t \ge 0$ including $t \to \infty$.

The construction of a sample path envelope for fBm traffic is, however, not straightforward since no simple overflow profile exists, see Sect. III. A simplifying approach is proposed in [29] where it is argued that the Weibull tail (4) implies an envelope for fBm traffic. Such an envelope is, however, based on the approximation by the largest term (3). A rigorous sample path envelope for fBm as well as end-to-end performance bounds under fBm cross traffic have not been derived.

III. GAMMA BOUND FOR FBM SAMPLE PATHS

In this section we derive a sample path envelope for fBm traffic. The envelope is an essential prerequisite for application of the stochastic network calculus. It is used to derive leftover service curves as well as end-to-end performance bounds in the following sections. In the remainder of this paper we consider a discrete time model, i.e. time t is a dimensionless counter of time slots each of fixed duration. Correspondingly, for the fBm traffic model (1) the rate λ is given in bits per time slot and the increment $Z(t)$ in bits. We use subscripted ε_p, ε_s, and ε_a to denote point-wise, sample path, and approximate or asymptotic overflow probabilities, respectively, see Sect. II.
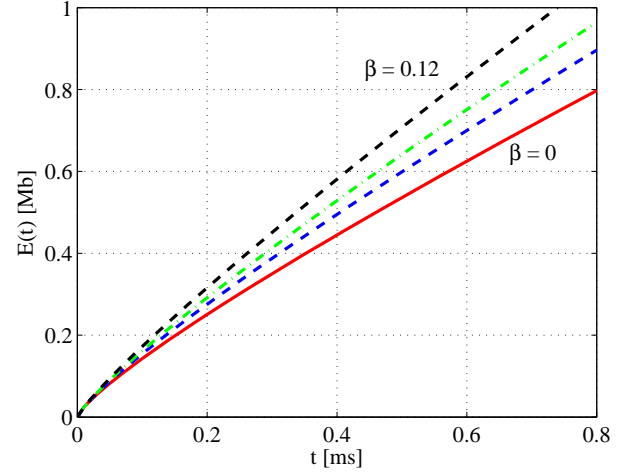
(a) The parameter β relaxes the envelope. The case $\beta = 0$ coincides with (6).
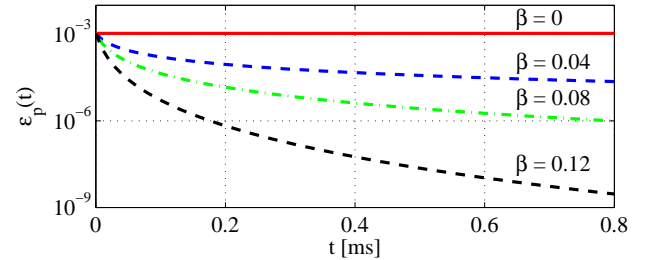
(b) The point-wise overflow probability $\varepsilon_p(t) = \eta^{t^{2\beta}}$ decays faster with increasing β but generally slower than exponentially since $\beta \in (0, 1 - H)$.

Fig. 1. FBm envelopes according to (9).

A. Sample Path Envelope

The difficulty of deriving a sample path envelope for fBm traffic is due to the intended integrability of the point-wise overflow probability $\int_0^\infty \varepsilon_p(t)dt$ as introduced in Sect. II. To this end, we consider the point-wise envelope (6) with a time-dependent overflow probability $\varepsilon_p(t)$

$$E(t) = \lambda t + \sqrt{-2\log \varepsilon_p(t)}\sigma t^H. \quad (9)$$

This envelope can be obtained from Chernoff's bound. It constitutes the natural shape of an fBm envelope in the sense that it is the optimal solution that can be derived in this way.

Here, we have to choose the point-wise overflow probability $\varepsilon_p(t)$ in such a way that it is integrable. In fact twofold integrability is required to apply network calculus composition results for tandem systems in Sect. V. At the same time $E(t)$ has to be at most linear in t for $t \to \infty$ to be able to derive performance bounds, e.g. $\limsup_{t \to \infty} E(t)/t < C$ for a constant rate server with capacity C. This constrains $\varepsilon_p(t)$ to decay no faster than proportionally to $e^{-t^{2-2H}}$ for $t \to \infty$. We make the practicable choice[3] $\varepsilon_p(t) = \eta^{t^{2\beta}}$ with

---

[3]We note that other choices can be made, e.g. an envelope with linear slack rate ϱ, i.e. $E(t) = (\lambda+\varrho)t+\sqrt{-2\log \eta}\sigma t^H$, follows at the cost of significant complexity if we set $\varepsilon_p(t) = \exp\left(-(\varrho t^{1-H} + \sqrt{-2\log \eta}\sigma)^2/(2\sigma^2)\right)$.

slack parameter $\beta \in (0, 1 - H)$ and $\eta \in (0, 1)$ to derive an envelope for fBm sample paths. The integrability of $\varepsilon_p(t)$ is established by Lem. 1 in the appendix. For the special case $\beta = 0$ the parameter η is the overflow probability of the point-wise envelope (6), i.e. $\varepsilon_p = \eta$ if $\beta = 0$.

Fig. 1 evaluates different fBm envelopes. The fBm traffic parameters are $\lambda = 0.5$, $\sigma = 0.5$, and $H = 0.7$, i.e. the traffic is LRD as frequently observed for the Internet [3], [4], [5], [6], [7]. To give the parameters a physical meaning we choose the time slot of the discrete time model to be 10 μs that is the transmission time of a 1250 Byte, respectively, 10 kb sized packet on a 1 Gb/s link. Accordingly, on a time slot basis λ is 5 kb that translates to 0.5 Gb/s.

Fig. 1(a) compares fBm envelopes (9) with $\varepsilon_p = \eta^{t^{2\beta}}$ that are slackened by $\beta = 0.04$, 0.08, and 0.12, respectively, to the envelope (6), i.e. the special case where $\beta = 0$. The corresponding point-wise overflow probability $\varepsilon_p(t)$ is shown in Fig. 1(b). Clearly, the overflow probability decays faster in case of larger β whereas it remains constant and hence non-integrable if $\beta = 0$. Generally, the overflow probability decays slower than exponentially for any choice of $\beta \in (0, 1 - H)$.

Theorem 1 (**FBM Sample Path Envelope**). *Given fBm traffic with mean rate λ, standard deviation σ, and LRD Hurst parameter $H \in (\frac{1}{2}, 1)$.*

$$E(t) = \lambda t + \sqrt{-2 \log \eta} \, \sigma t^{H + \beta}$$

satisfies the definition of sample path envelope (7) with overflow probability

$$\varepsilon_s = \frac{\Gamma(\frac{1}{2\beta})}{2\beta(-\log \eta)^{\frac{1}{2\beta}}}$$

where $\beta \in (0, 1 - H)$ and $\eta \in (0, 1)$ are free parameters.

Proof: Assuming stationarity of the arrivals and letting $t \to \infty$ the sample path envelope (7) can be written as

$$\mathsf{P}\left[\sup_{\tau \geq 0}\{A(\tau) - E(\tau)\} > 0\right] \leq \varepsilon_s. \quad (10)$$

The overflow probability is defined by a union of events

$$\mathsf{P}\left[\sup_{\tau \geq 0}\{A(\tau) - E(\tau)\} > 0\right] = \mathsf{P}\left[\bigcup_{\tau = 0}^{\infty}\{A(\tau) - E(\tau) > 0\}\right]$$

that can be estimated by application of Boole's inequality

$$\mathsf{P}\left[\bigcup_{\tau = 0}^{\infty}\{A(\tau) > E(\tau)\}\right] \leq \sum_{\tau = 1}^{\infty} \mathsf{P}[A(\tau) > E(\tau)]$$
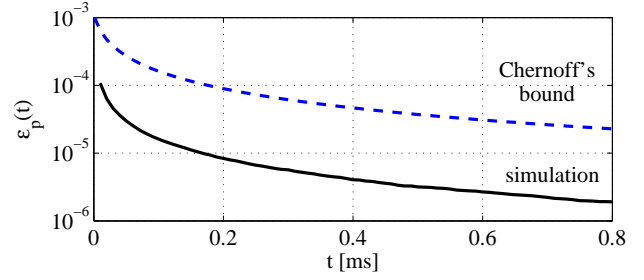
where we used the fact that the overflow probability at $\tau = 0$ is trivially zero since by definition $A(0) = E(0) = 0$.

Applying Chernoff's bound and using the effective bandwidth of fBm (5) we have for any $\theta > 0$ that
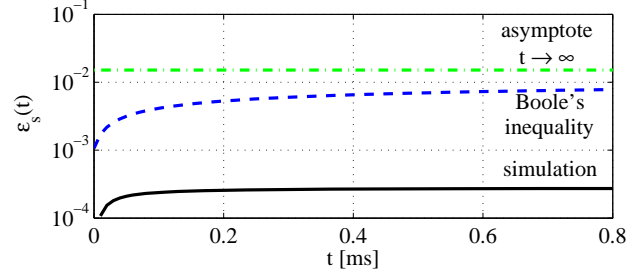
$$\mathsf{P}[A(\tau) > E(\tau)] \leq e^{-\theta E(\tau)} \mathsf{E}\left[e^{\theta A(\tau)}\right] = e^{-\theta E(\tau)} e^{\theta \lambda \tau + \frac{\theta^2 \sigma^2}{2} \tau^{2H}}.$$

Inserting $E(\tau) = \lambda \tau + \sqrt{-2 \log \eta} \, \sigma \tau^{H + \beta}$ and minimizing over $\theta > 0$ yields $\theta = \frac{1}{\sigma}\sqrt{-2 \log \eta} \, \tau^{\beta - H}$ and by insertion

$$\mathsf{P}[A(\tau) > E(\tau)] \leq \eta^{\tau^{2\beta}}.$$

(a) Point-wise overflow probability $\varepsilon_p(t) = \eta^{t^{2\beta}}$ vs. simulation results. The overflow probabilities are conservative due to the use of Chernoff's bound.

(b) Sample path overflow probability ε_s vs. simulation results for sample paths of length t. The overflow probabilities are strict upper bounds due to the sample path argument using Boole's inequality. For $t \to \infty$ the sample path overflow probability stays finite and converges to the result obtained from Th. 1.

Fig. 2. Overflow probabilities for envelopes from Th. 1 where $\beta = 0.04$.

Summing the point-wise overflow probabilities gives

$$\sum_{\tau = 1}^{\infty} \eta^{\tau^{2\beta}} \leq \int_0^{\infty} \eta^{\tau^{2\beta}} d\tau = \frac{\Gamma(\frac{1}{2\beta})}{2\beta(-\log \eta)^{\frac{1}{2\beta}}}$$

where we used that $\eta^{\tau^{2\beta}}$ is monotonically decreasing in τ to estimate each summand indexed by τ by an integral over $(\tau - 1, \tau]$. Finally, we applied Lem. 1 from the appendix. ∎

Fig. 2(a) compares the point-wise overflow probability $\varepsilon_p(t) = \eta^{t^{2\beta}}$ of the envelope from Fig. 1(a) for $\beta = 0.04$ with simulation results obtained from 10^9 fBm sample paths generated with Matlab. The overflow probabilities are computed from Chernoff's bound and hence are conservative. Fig. 2(b) shows the overflow probability for sample paths of length t. The simulation results are obtained by counting the number of sample paths that violate the envelope at least once in $[0, t]$. To derive the analytical bound we sum the point-wise overflow probabilities over all $\tau \in [0, t]$. The dashed horizontal line shows the overflow probability of the sample path envelope from Th. 1 that holds for sample paths of any length $t \geq 0$, i.e. the sum of the point-wise overflow probabilities stays finite for all t and converges to the dashed line for $t \to \infty$.

B. Performance Bounds

We use the sample path envelope from Th. 1 to derive performance bounds for a server fed with fBm traffic. Owing to a known duality between backlog bounds and envelopes we find affine envelopes for fBm traffic along the way, see Sect. III-D.

Theorem 2 (**Backlog and Delay Bound**). *Consider a lossless work-conserving constant rate server with capacity C fed with fBm traffic as in Th. 1 where $C > \lambda$. The steady state backlog B is bounded by b subject to the overflow probability ε_s*

$$\mathsf{P}[B > b] \leq \varepsilon_s = \frac{\Gamma(\frac{1}{2\beta})}{2\beta(-\log \eta)^{\frac{1}{2\beta}}}$$

where $\beta \in (0, 1 - H)$ is a free parameter and

$$\eta = \exp\left(-\frac{1}{2\sigma^2}\left(\frac{C - \lambda}{H + \beta}\right)^{2(H+\beta)}\left(\frac{b}{1 - (H + \beta)}\right)^{2-2(H+\beta)}\right).$$

The steady state delay under first-come first-serve (fcfs) scheduling W is bounded by $\mathsf{P}[W > b/C] \leq \varepsilon_s$.

Proof: Assuming stationarity and letting $t \to \infty$ we obtain the steady state backlog from (2) as

$$\mathsf{P}[B > b] = \mathsf{P}\left[\sup_{\tau \geq 0}\{A(\tau) - C\tau\} > b\right].$$

The expression is a special case of sample path envelope (10). Define $E(t)$ as in Th. 1. If $E(t) \leq b + Ct$ for all $t \geq 0$ then

$$\mathsf{P}[B > b] \leq \mathsf{P}\left[\sup_{\tau \geq 0}\{A(\tau) - E(\tau)\} > 0\right] \leq \frac{\Gamma(\frac{1}{2\beta})}{2\beta(-\log \eta)^{\frac{1}{2\beta}}}.$$

Given b and C we derive the largest envelope that satisfies the constraint $E(t) \leq b + Ct$ for all $t \geq 0$. To this end, we first find $t = \tau^*$ that minimizes the vertical distance between $E(t)$ and $b + Ct$. Hence, τ^* is the solution of $\partial E(t)/\partial t = C$. Then, we choose the parameter $\eta \in (0, 1)$ such that $E(t)$ and $b + Ct$ are tangent to each other at τ^*, i.e. $E(\tau^*) = b + C\tau^*$. From the first requirement we derive τ^* as

$$\tau^* = \left(\frac{\sqrt{-2\log\eta}\,\sigma(H + \beta)}{C - \lambda}\right)^{\frac{1}{1-(H+\beta)}}. \qquad (11)$$

By insertion of τ^* into the condition $E(\tau) = b + C\tau$ the optimal parameter η follows as given in Th. 2.

The delay bound can be derived as the maximum horizontal distance of $E(t)$ and Ct using the same basic steps. ∎

An approach related to Th. 2 has been employed in [34] using, however, the point-wise fBm envelope (6) and the principle of the largest term (3) to compute an approximate backlog bound. In contrast, our sample path envelope does not resort to this approximation and establishes rigorous upper bounds.

We recover previous results for the special case where $\beta = 0$. In this case the optimal parameter η derived in Th. 2 is equal to the overflow probability of the known asymptotic backlog bound (4), i.e. $\eta = \varepsilon_a$. Recall that (4) has been derived using either the Gaussian distribution of the fBm increments [10], [11], large deviations theory [12], or the point-wise envelope (6) [34]. All these derivations are based on the approximation by the largest term (3). Similarly, η is the point-wise overflow probability of the fBm envelope $E(t)$ from (6), i.e. $\eta = \varepsilon_p$, where $E(t)$ is the largest envelope that is smaller than $b + Ct$ for all $t \geq 0$. Furthermore, $E(t)$ is tangent at τ^* such that
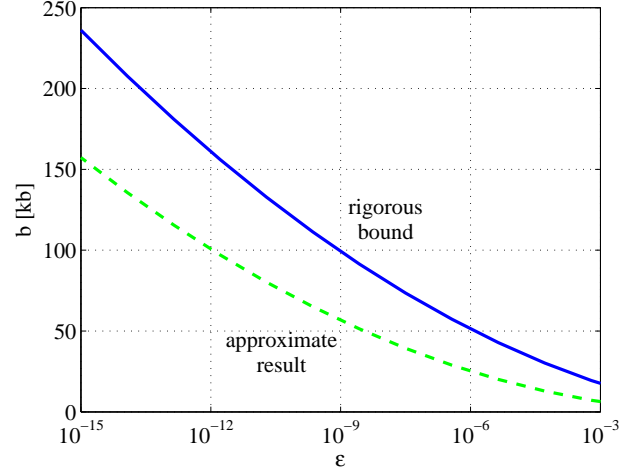

Fig. 3. Rigorous backlog bound from Th. 2 in comparison to the approximation resp. asymptotic result from (4) for fBm traffic with parameters $\lambda = 0.5$ Gb/s, $\sigma = 0.25$ Gb/s, and $H = 0.75$ at a server with capacity $C = 1$ Gb/s. Both bounds decay slower than exponentially due to LRD.

$b + Ct$ generally has an overflow probability smaller than η except at τ^* where it attains its maximum that equals η. Hence, the supremum in (3) is attained at τ^* and the boundary point at τ^* is the most probable point for violation of $b + Ct$, i.e. for buffer overflow. From (11) the most probable time scale for violation evaluates by insertion of η from Th. 2 and $\beta = 0$ to

$$\tau^* = \left(\frac{b}{C - \lambda}\right)\left(\frac{H}{1 - H}\right)$$

recovering the result from [10]. Related probabilistic bounds for the duration of busy periods τ' at a constant rate server fed with fBm traffic have been derived in [33] as the point in time where the point-wise envelope (6) equals Ct

$$\tau' = \left(\frac{\sqrt{-2\log\varepsilon_p}\,\sigma}{C - \lambda}\right)^{\frac{1}{1-H}}.$$

It is interesting to observe the large impact of H on the relevant time scales that is unfavorable for the use time scale bounds.

Fig. 3 displays backlog bounds for a server with capacity $C = 1$ Gb/s fed with fBm traffic with parameters $\lambda = 0.5$ Gb/s, $\sigma = 0.25$ Gb/s, and $H = 0.75$. The parameter set agrees with observations in [11] and will be used in the sequel unless mentioned otherwise. The overflow probabilities are obtained from Th. 2 using sample path arguments (solid line) and from (4) using the principle of the largest term (dashed line), respectively. We optimize parameter β numerically. Both results agree with each other regarding the slower than exponential decay that is due to LRD. The absolute values differ, however, by a factor of about 1.75 at $\varepsilon = 10^{-9}$ due to the fact that Th. 2 provides a rigorous upper bound whereas (4) is an approximation. The relative difference of the two results decreases for smaller ε.

Fig. 4 shows how a variation of the different parameters of the fBm traffic affects the backlog bounds. Clearly, for larger

(a) Impact of mean rate λ

(b) Impact of standard deviation σ

(c) Impact of H, moderate H
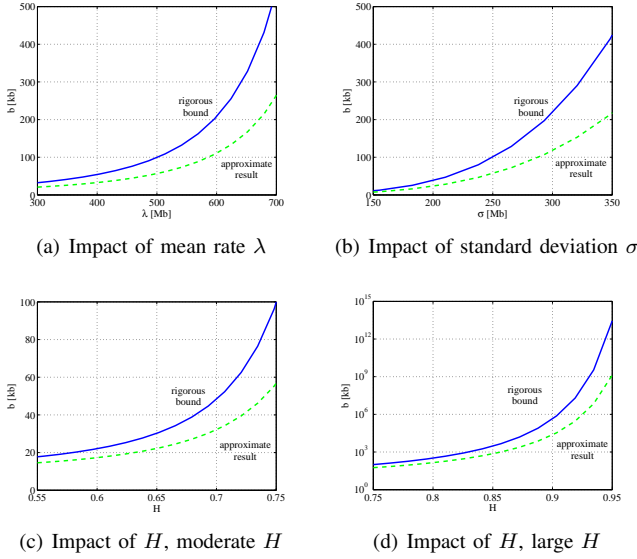

(d) Impact of H, large H

Fig. 4. Effect of different traffic parameters on backlog bounds as in Fig. 3 with overflow probability $\varepsilon = 10^{-9}$. Note the tremendous impact of the Hurst parameter, hence Fig. 4(d) is on log scale.

λ, σ, or H both backlog bounds increase. For moderate H the rigorous and the approximate bounds agree well, whereas large H have a huge impact, hence Fig. 4(d) is on log scale.

Fig. 4(b) can also be interpreted as showing the effect of statistical multiplexing of fBm flows. Given m statistically independent fBm flows each normalized by m, i.e. with rate λ/m and standard deviation σ/m and identical Hurst parameter, the aggregate traffic of all m flows has rate λ but standard deviation $\sigma/\sqrt{m}$. Statistical multiplexing of fBm flows is discussed in detail in [34].

C. Weibullian Decay of Overflow Probabilities

In this subsection we show that overflow probabilities of fBm sample path bounds have a Weibull tail in b, i.e. $\log \varepsilon_s \sim -b^{2-2H}$. To this end, we minimize ε_s from Th. 2 over β.

To obtain the results shown in Fig. 3 and Fig. 4 from Th. 2 we optimized the free parameter $\beta \in (0, 1-H)$ numerically. We note that the optimal choice of β has significant impact, e.g. on the decay of ε with b. Typically, we find that the optimal β is small compared to one. In this case a useful approximation of the Gamma bound in Th. 1 and Th. 2 is

$$\tilde{\varepsilon}_s = \frac{\sqrt{\pi}}{\sqrt{\beta}(2e\beta(-\log \eta))^{\frac{1}{2\beta}}} \quad \text{for} \quad \beta \ll 1 \qquad (12)$$

using Stirling's formula $\Gamma(x) \approx \sqrt{2\pi/x}\,(x/e)^x$ for $x \gg 1$. The approximation (12) is exact in the limit $\beta \to 0$.

In the remainder of this subsection we optimize the free parameter $\beta \in (0, 1-H)$. We use (12) and minor simplifications for small β to derive a near optimal solution β^*. We emphasize that Th. 2 holds for any $\beta \in (0, 1-H)$ where we approximate β^* that minimizes the bound, i.e. inserting β^* into Th. 2 yields a rigorous upper bound that is close to its minimal solution.

Assuming $\beta \ll (1-H)$ and approximating $1 - H - \beta$ by $1-H$ we compute the derivative of (12) and solve $\partial\tilde{\varepsilon}_s/\partial\beta = 0$ for β. we find that the minimum of (12) is approached at $\beta = -W(1/(2\log \varepsilon_a))$ where ε_a coincides with (4). $W(z)$ denotes Lambert's W function that is the inverse of $z = xe^x$. It is real-valued for $\varepsilon_a < e^{-e/2}$. Since β is assumed to be small a good approximation of the optimal solution is

$$\beta^* = \frac{1}{2(-\log \varepsilon_a)}$$

where we estimate the Lambert W function by a linear segment. Note that β^* decreases with b. We define the quotient

$$\chi = \left(\frac{H^H(1-H)^{1-H}}{(H+\beta)^{H+\beta}(1-(H+\beta))^{1-(H+\beta)}} \right)^2 \qquad (13)$$

that is in $[\frac{1}{4}, 1]$ for $H \in (\frac{1}{2}, 1)$, $\beta \in (0, 1-H)$ and approaches 1 for small β. Inserting β^* into (12) we find the closed form

$$\tilde{\varepsilon}_s = \frac{b}{C - \lambda} \varepsilon_a^{1+\log \chi} \sqrt{2\pi(-\log \varepsilon_a)} \qquad (14)$$

as a near optimal solution of Th. 2. We note that (14) matches the numerically optimized results in Fig. 3 and 4 almost perfectly which is why we omit reproducing similar graphs.

Phrasing ε_s as a function of ε_a enables us to support significant conclusions on the dimensioning of networks from the largest term approximation by sample path arguments. An important example is the relation of spare capacity $C - \lambda$ and buffer size b. While for $H = \frac{1}{2}$ halving $C - \lambda$ requires doubling b to achieve constant ε_a (4) the tradeoff deteriorates for large H [11]. Under LRD spare capacity becomes more important and buffering much less efficient supporting current arguments for reducing router buffers [2].

Inserting ε_a from (4) into (14), bounding χ from below by a constant that is close to 1 for $\beta \ll 1-H$, and using positive constants c_i yields after simplification that

$$\log \tilde{\varepsilon}_s = -c_1 b^{2-2H} + \log(c_2 b^{2-H}) + c_3.$$

Based on sample path arguments the expression reflects the decay rate of the largest term approximation (4).

To illustrate the tail behavior we compare the overflow probability of backlog bounds for fBm traffic to traffic with exponentially bounded burstiness (EBB) [26]. For generation of EBB traffic we use a discrete time Markov model with two states: on and off. In the off state (state 1) no traffic is generated and in the on state (state 2) traffic is generated with peak rate P. The steady state probability of the on state is $p_{\text{on}} = p_{12}/(p_{12} + p_{21})$ where p_{ij} are the state transition probabilities from state i to state j. The mean rate becomes $\lambda = p_{\text{on}} P$. As [19] we characterize the burstiness of an on-off source by the average time for two state changes to occur $T = 1/p_{12} + 1/p_{21}$. The aggregate of m on-off sources has a point-wise envelope (8) given as $E(t) = m\rho(\theta)t$ with overflow profile $\varepsilon_p(b) = e^{-\theta b}$, see [19]. The envelope rate $\rho(\theta)$ of a single discrete time on-off source is given in [15] as

$$\frac{1}{\theta}\log\left(\frac{p_{11}+p_{22}e^{\theta P}+\sqrt{(p_{11}+p_{22}e^{\theta P})^2-4(p_{11}+p_{22}-1)e^{\theta P}}}{2} \right)$$

TABLE I
TRAFFIC PARAMETERS USED FOR FIG. 5.

type	# sources	mean rate	variability	correlation
EBB	$m = 100$	$\lambda = 5$ Mb/s	$P = 5\lambda$	T
fBm	$m = 1$	$\lambda = 500$ Mb/s	$\sigma = \lambda/2$	H

for $\theta > 0$ where $\rho(\theta)$ ranges between the mean and the peak rate of the source. Employing the envelope a backlog bound follows from (2) by application of Boole's inequality and integration of the overflow profile, see Sect. II and [19] for details. Given aggregate traffic generated by m on-off sources fed into a server with capacity C the overflow probability $\mathsf{P}[B > b] \leq \varepsilon_s$ of the backlog bound b becomes

$$\varepsilon_s = \int_0^\infty e^{-\theta(b+(C-m\rho(\theta))\tau)} d\tau = \frac{e^{-\theta b}}{\theta(C - m\rho(\theta))} \quad (15)$$

for any $\theta > 0$ that satisfies $m\rho(\theta) < C$.

For Fig. 5 we employ the traffic parameters given in Tab. I. We use a traffic aggregate consisting of 100 EBB flows. In case of fBm we use only a single flow. Recall that fBm is typically used as a model for aggregate traffic, e.g. it has been related to the superposition of on-off sources with heavy-tailed on and off periods [8], [5], [6]. In contrast the on-off sources used for EBB traffic have geometrically distributed on and off periods. The fBm traffic has the same parameters as used for Fig. 3 and Fig. 4.

Fig. 5 clearly shows the Weibullian decay of the fBm backlog bound (14) with $\log \varepsilon_s \sim -b^{2-2H}$ that becomes slower with increasing H, i.e. with more pronounced LRD even very large buffers are filled with a certain probability. In contrast, the decay is exponentially fast in case of EBB traffic where the correlation parameter T affects only the slope. From Fig. 5 we conclude that reasonable performance guarantees for EBB as well as fBm traffic exist if a certain violation probability is tolerated. A near-deterministic service with very small violation probability is, however, only practicable for EBB traffic. Due to the slow tail decay it is very costly to reduce the overflow probability for fBm traffic with LRD. We mention that the approximation (14) as before agrees well with Th. 2. For EBB we optimized the parameter θ numerically.

D. Affine Envelopes

We conclude this section with a corollary on affine fBm envelopes that follow from the backlog bounds. The proof of Th. 2 already shows a close link between arrival envelopes and backlog bounds. This relation has been elaborated in [29] where a traffic model referred to as generalized Stochastically Bounded Burstiness (gSBB) is defined. The gSBB traffic characterization uses an affine sample path envelope with defined overflow profile. A general definition of sample-path envelope with overflow profile $\varepsilon(b)$ [24], [19] extends (7) to

$$\mathsf{P}\left[\sup_{\tau \in [0,t]} \{A(\tau, t) - E(t - \tau)\} > b \right] \leq \varepsilon_s(b). \quad (16)$$

The gSBB envelope is the special case where $E(t) = rt$. The finding in [29] is that given a server with capacity C and
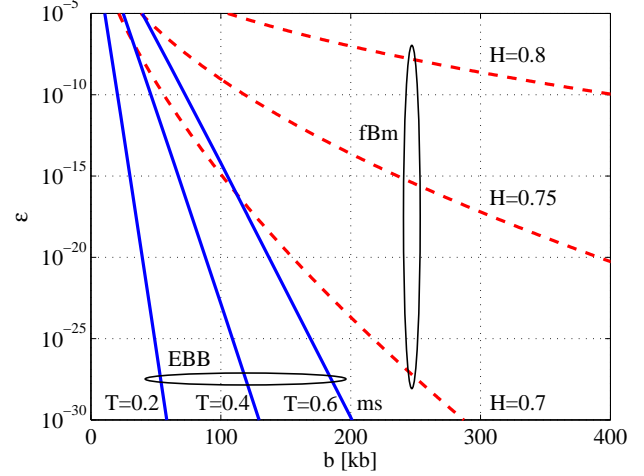

Fig. 5. Weibullian decay of the overflow probability for fBm traffic where $\log \varepsilon \sim -b^{2-2H}$ compared to the exponential decay for EBB traffic.

statistical backlog bound b with overflow probability $\mathsf{P}[B > b] \leq \varepsilon(b)$ the arrivals of the system are gSBB and satisfy the definition of sample path envelope (16) with $E(t) = Ct$ and identical overflow profile $\varepsilon(b)$. It is argued in [29] that a gSBB envelope for fBm traffic follows immediately from (4).

Corollary 1 (**Affine FBM Envelopes**). *Given fBm traffic $E(t) = rt$ is a sample path envelope (16) with overflow profile*

$$\varepsilon_s(b) = \frac{\Gamma(\frac{1}{2\beta})}{2\beta\vartheta^{\frac{1}{2\beta}}} b^{-\frac{1-(H+\beta)}{\beta}}$$

where $\beta \in (0, 1-H)$ is a free parameter and

$$\vartheta = \frac{1}{2\sigma^2}\left(\frac{r-\lambda}{H+\beta}\right)^{2(H+\beta)}\left(\frac{1}{1-(H+\beta)}\right)^{2-2(H+\beta)}.$$

An approximate overflow profile is $\varepsilon_a(b) = e^{-\upsilon b^{2-2H}}$ where

$$\upsilon = \frac{1}{2\sigma^2}\left(\frac{r-\lambda}{H}\right)^{2H}\left(\frac{1}{1-H}\right)^{2-2H}.$$

The first statement of the corollary uses our backlog bound from Th. 2 that is based on the sample path envelope from Th. 1. The second statement uses (4) as proposed in [29]. Note, however, that (4) is based on the approximation by the largest term (3), i.e. it does not provide a sample path envelope.

The derived sample path envelopes are an essential building block for application of the stochastic network calculus. In Sect. IV we will use the envelopes to derive a leftover service curve to analyze systems where fBm cross traffic is multiplexed, scheduled, and de-multiplexed afterwards. In Sect. V we will compose these service curves by convolution to explore tandem systems, each under fBm cross traffic. We will use the affine envelopes established by Cor. 1. The envelope from Th. 1 can be used in the same way and may yield tighter bounds at the cost of additional complexity.

TABLE II
TRAFFIC PARAMETERS USED FOR FIG. 6, FIG. 8, AND FIG. 9.

traffic	type	# sources	mean rate	variability	correlation
through	CBR	$m = 100$	$\lambda = 2.5$ Mb/s	–	–
	EBB	$m = 100$	$\lambda = 2.5$ Mb/s	$P = 5\lambda$	T^{th}
	fBm	$m = 1$	$\lambda = 250$ Mb/s	$\sigma = \lambda/2$	H^{th}
cross	EBB	$m = 100$	$\lambda = 2.5$ Mb/s	$P = 5\lambda$	T^{cr}
	fBm	$m = 1$	$\lambda = 250$ Mb/s	$\sigma = \lambda/2$	H^{cr}

IV. LEFTOVER SERVICE UNDER FBM CROSS TRAFFIC

The network calculus uses the concept of service curves $S(t)$ to model the service provided by a system. The service curve relates a system's departures $D(t)$ to its arrivals $A(t)$ [23]. A fundamental stochastic service curve is defined in [19] and likewise in [24]

$$\mathsf{P}\left[D(t) < \inf_{\tau \in [0,t]}\left\{A(\tau) + [S(t-\tau) - b]_+\right\}\right] \leq \varepsilon(b) \quad (17)$$

where $[x]_+ = \max\{0, x\}$. The stochastic service curve is subject to a deficit profile $\varepsilon(b)$ that is decreasing in b. The inner operation $\inf_{\tau \in [0,t]}\{f(\tau) + g(t-\tau)\} := f \otimes g(t)$ is referred to as the convolution under the min-plus algebra.

So-called leftover service curves can effectively characterize the service that remains for a through flow at a system after scheduling cross traffic [28], [19]. A basic leftover service curve that does not make any assumptions about the order of scheduling can be deduced by subtracting the sample path envelope of the cross traffic from the service provided by the system. Given a server with capacity C and cross traffic with sample path envelope $E(t)$ and overflow profile $\varepsilon_s(b)$ as in (16). A leftover service curve that satisfies (17) with deficit profile $\varepsilon_s(b)$ is $S(t) = Ct - E(t)$. The following corollary uses the affine fBm sample path envelope from Cor. 1 to characterize the leftover service under fBm cross traffic.

Corollary 2 (**FBM Leftover Service Curve**). *Consider a server with capacity C under fBm cross traffic. A service curve (17) for through traffic is $S(t) = (C - r)t$ with deficit profile $\varepsilon(b)$. The deficit profile equals the overflow profile $\varepsilon_s(b)$ of the fBm envelope in Cor. 1. It is approximated by $\varepsilon_a(b)$, Cor. 1.*

The definitions of service curve and sample path arrival envelope facilitate the derivation of performance bounds. Given arrivals with sample path envelope $E(t)$ (16) at a system with service curve $S(t)$ (17), the steady state virtual backlog B is stochastically bounded by [24], [19]

$$\mathsf{P}\left[B > \sup_{\tau \geq 0}\{E(\tau) - S(\tau)\} + b\right] \leq \varepsilon(b)$$

where $\varepsilon(b) = \varepsilon^{th} \otimes \varepsilon^{cr}(b)$. We use superscripted ε^{th} and ε^{cr} to distinguish overflow profiles of through and cross traffic. The latter matches the deficit profile of the service curve. The fcfs waiting time W is bounded by $\mathsf{P}[W > w] \leq \varepsilon(b)$ where

$$w = \inf\{\tau \geq 0 : S(t+\tau) \geq E(t) + b \ \forall t \geq 0\}. \quad (18)$$

Backlog and delay can be visualized as the vertical, respectively, horizontal deviation of the arrival envelope and the service curve subject to the overflow and deficit profiles.
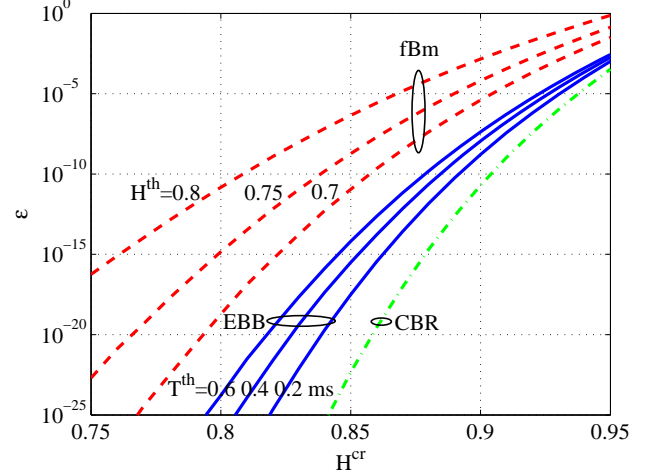

Fig. 6. Violation probability of a delay bound of 1 ms for different types of through traffic at a constant rate server with fBm cross traffic. The Hurst parameter of the cross traffic H^{cr} has significant impact on the performance of the through traffic. While the burstiness of through traffic is crucial if H^{cr} is small, the cross traffic dominates if H^{cr} is large.

We derive performance bounds for a through flow under fBm cross traffic using the leftover service curve from Cor. 2. We consider three fundamentally different types of through traffic: (*a*) constant bit rate (CBR), (*b*) traffic with exponentially bounded burstiness (EBB) [26], and (*c*) fBm with LRD, i.e. slower than exponential burstiness decay.

All three types of through traffic are configured to have the same mean rate λ. We generally use affine sample path envelopes (16). Trivially, the envelope of the CBR traffic is $E(t) = rt$ with $r = \lambda$ and overflow profile $\varepsilon_s(b) = 0$ for all $b \geq 0$. For generation of EBB traffic we use the discrete time Markov model introduced in Sect. III-C. A sample path envelope (16) for the aggregate traffic generated by m on-off sources follows from the backlog bound (15) as $E(t) = rt$ with overflow profile $\varepsilon_s(b) = e^{-\theta b}/(\theta(r - m\rho(\theta)))$ for any $\theta > 0$ that satisfies $m\rho(\theta) < r$. For the fBm through traffic we use the envelope from Cor. 1.

The delay bound follows from (18) for $r^{th} + r^{cr} \leq C$ as

$$\mathsf{P}\left[W > \frac{b^{th} + b^{cr}}{C - r^{cr}}\right] \leq \varepsilon^{th}(b^{th}) + \varepsilon^{cr}(b^{cr}).$$

Here, b^{th}, r^{th} are the parameters of the through traffic envelope and b^{cr}, r^{cr} are the parameters of the cross traffic envelope that determine the leftover service curve. Note that different sets of parameters can yield different violation probabilities for the same delay bound. To obtain the best possible result we optimize the parameters of the envelopes including the parameter β for each traffic type numerically.

In Fig. 6 we display the violation probability of a delay bound of 1 ms for CBR, EBB, and fBm through traffic at a server with capacity $C = 1$ Gb/s under fBm cross traffic. Through and cross traffic are configured to have the same mean rate. The traffic parameters are summarized in Tab. II. Fig. 6 shows the huge impact of the Hurst parameter H^{cr} of the cross

traffic on the performance of through flows. While type and burstiness of the through traffic largely determine the violation probability of the delay bound, the influence becomes much less pronounced for cross traffic with large H^{cr}, i.e. the LRD of the cross traffic becomes the dominating effect.

V. END-TO-END PERFORMANCE BOUNDS

The particular strength of the network calculus is its ability to characterize the end-to-end service of tandem systems. The service curves of individual systems can be composed by min-plus convolution into a network service curve. This network service curve models a whole network as if it were a single system. Hence, it facilitates the derivation of end-to-end performance bounds using single system results, e.g. (18).

A. Sample Path FBM Leftover Service Curve

The easy composition of tandem systems in the network calculus is due to the associativity of min-plus convolution. For the special case of deterministic systems (17) has overflow profile $\varepsilon(b) = 0$ for all $b \geq 0$ and reduces to $D(t) \geq A \otimes S(t)$ almost surely. Given a network of two systems with deterministic service curves $S^1(t)$ and $S^2(t)$ in series the departures of the first system $D^1(t) = A^1 \otimes S^1(t)$ are the arrivals to the second system $A^2(t) = D^1(t)$ and by recursive insertion it follows that $D^2(t) \geq A^2 \otimes S^2(t) = (A^1 \otimes S^1) \otimes S^2(t)$. Using associativity of min-plus convolution $D^2 \geq A^1 \otimes S^{net}(t)$ where $S^{net} = S^1 \otimes S^2(t)$ is the network service curve.

The composition of stochastic service curves is, however, much more involved. The difficulty is due to the fact that a recursive insertion of the stochastic service curve (17) is not possible since (17) uses sample paths of the arrivals $A(t)$ but makes only a point-wise statement for the departures $D(t)$. To derive stochastic network service curves an extended definition of stochastic service curve that makes sample path guarantees for the departures is required. The problem does not occur in the deterministic case, see [19], [35] for details.

A fundamental stochastic network service curve is derived in [19], [35]. We make a marginal adaptation of the service curve to discrete time that is used in this work. For n systems in series each with service curve $S^i(t)$ and deficit profile $\varepsilon^i(b)$ according to (17) a network service curve is

$$S^{net}(t) = S^1 \otimes S_{-\delta}^2 \otimes \cdots \otimes S_{-(n-1)\delta}^n(t) \quad (19)$$

where $S_{-\delta}(t) = S(t) - \delta t$ and $\delta > 0$ is a free parameter. The network service curve satisfies (17) with deficit profile

$$\varepsilon^{net}(b) = \varepsilon_\delta^1 \otimes \varepsilon_\delta^2 \otimes \cdots \otimes \varepsilon_\delta^{n-1} \otimes \varepsilon^n(b). \quad (20)$$

The deficit profiles $\varepsilon_\delta^i(b)$ stem from an extended definition of stochastic service curve that makes guarantees for entire sample paths of the departures $D(t)$ as opposed to (17) that only makes a point-wise statement. To derive such sample path guarantees [19], [35] contributes an essential sample path service curve that is relaxed by parameter $\delta > 0$. The definition of sample path service curve states that

$$P\left[\sup_{t \in [0,u]}\left\{\inf_{\tau \in [0,t]}\left\{A(\tau) + [S(t-\tau) - \delta(u-t) - b]_+\right\} - D(t)\right\} > 0\right]$$
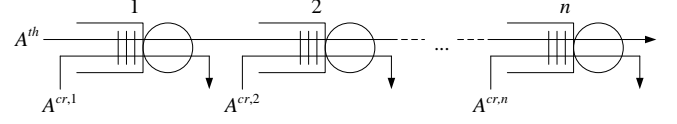

Fig. 7. A through flow traverses a tandem of n queuing systems. Cross traffic is multiplexed and de-multiplexed at each system

is smaller equal $\varepsilon_\delta(b)$. The relaxation by δ permits deriving the sample path deficit profile from the point-wise deficit profile of the departures using Boole's inequality as [19], [35]

$$\varepsilon_\delta(b) = \frac{1}{\delta}\int_b^\infty \varepsilon(x)dx. \quad (21)$$

We use the concept of network service curve to analyze tandem systems where fBm cross traffic is multiplexed and de-multiplexed. To this end, we derive the sample path deficit profile (21) for a leftover service curve under fBm cross traffic.

Corollary 3 (Sample Path FBM Leftover Service Curve). *Consider a server with capacity C under fBm cross traffic as given in Cor. 1. The leftover service curve $S(t) = (C-r)t$ relaxed by rate $\delta > 0$ has the sample path deficit profile (21)*

$$\varepsilon_\delta(b) = \frac{\Gamma(\frac{1}{2\beta})}{2\delta\vartheta^{\frac{1}{2\beta}}(1-(H+2\beta))} b^{-\frac{1-(H+2\beta)}{\beta}}$$

where $\beta \in \left(0, \frac{1-H}{2}\right)$ is a free parameter and ϑ from Cor. 1.

Cor. 3 follows by insertion of the deficit profile from Cor. 2, respectively, Cor. 1 into (21) as

$$\varepsilon_\delta(b) = \frac{\Gamma(\frac{1}{2\beta})}{2\beta\delta\vartheta^{\frac{1}{2\beta}}}\int_b^\infty x^{-\frac{1-(H+\beta)}{\beta}}dx$$

that has a finite solution if $\beta < \frac{1-H}{2}$. Alternatively, the approximate overflow profile $\varepsilon_a(b)$ from Cor. 1 can be used to derive a similar result where b^{2-2H} becomes, however, the second argument of an incomplete Gamma function.

B. Scaling of End-to-end Performance Bounds

Cor. 3 enables the derivation of end-to-end service curves for networks under fBm cross traffic. In the remainder of this section we consider the line topology shown in Fig. 7 where fBm cross traffic is multiplexed and de-multiplexed at each hop. For ease of notation we assume n homogeneous systems in series, each with capacity C and fBm cross traffic with identical parameters λ, σ, and H^{cr}. We show that end-to-end performance bounds for n tandem systems under fBm cross traffic grow in $\mathcal{O}\left(n(\log n)^{\frac{1}{2-2H}}\right)$.

The leftover service curve at each system is $S^i(t) = (C - r^{cr})t$ where r^{cr} is the envelope rate of the fBm cross traffic. From (19) we obtain $S^{net}(t) = (C - r^{cr} - \Delta)t$ where $\Delta = (n-1)\delta$ is in $(0, C - r^{cr})$ and used as a constant. The deficit profile follows from (20) by insertion of Cor. 2 and Cor. 3 as

$$\varepsilon^{net}(b) = \frac{\Gamma(\frac{1}{2\beta})}{2\vartheta^{\frac{1}{2\beta}}}\inf_x\left\{\frac{(n-1)^2(\frac{b-x}{n-1})^{-\frac{1-(H+2\beta)}{\beta}}}{\Delta(1-(H+2\beta))} + \frac{x^{-\frac{1-(H+\beta)}{\beta}}}{\beta}\right\}$$
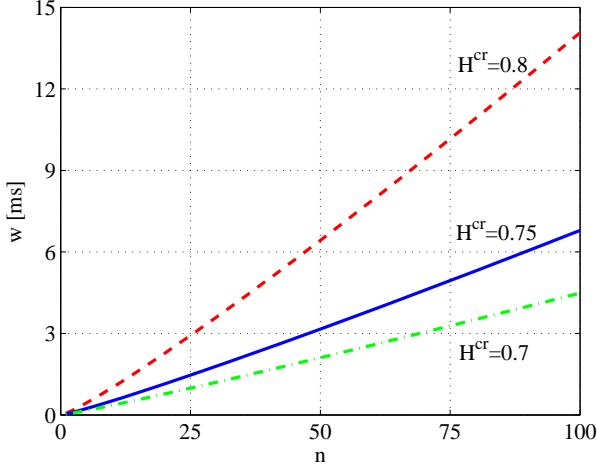
Fig. 8. End-to-end delay bounds with violation probability $\varepsilon = 10^{-12}$ for n constant rate servers with fBm cross traffic in series. The delay bounds shown for different H^{cr} of cross traffic grow slightly faster than linearly.

where $x \in (0, b)$ and $\beta \in \left(0, \frac{1-H}{2}\right)$.

Fig. 8 shows end-to-end delay bounds for a CBR through flow that traverses n tandem systems each with capacity C and fBm cross traffic. The delay bound is computed from (18) using the network service curve and CBR through traffic as $\mathsf{P}[W > b/(C - r^{cr} - \Delta)] \leq \varepsilon^{net}(b)$ under the constraint that $r^{th} + r^{cr} + \Delta \leq C$. The parameters of through and cross traffic are given in Tab. II. As before, we optimize the free parameters of the envelopes numerically.

Fig. 8 once more confirms the huge impact of cross traffic with LRD on network performance. After numerical optimization of β Fig. 8 shows a slightly faster than linear growth of end-to-end delay bounds with the number of tandem systems n subject to a violation probability $\varepsilon = 10^{-12}$.

We derive an analytical result on the scaling of end-to-end delay bounds with the number of systems in series using Stirling's formula as in Sect. III-C. The growth of end-to-end delay bounds with n is determined by the tail decay of the fBm overflow, respectively, deficit profiles. We neglect the irregularity in (20) that is due to the last hop and estimate ε^n by ε_δ^n to obtain the simplified deficit profile

$$\varepsilon^{net}(b) = \frac{\Gamma(\frac{1}{2\beta})n(n-1)(\frac{b}{n})^{-\frac{1-(H+2\beta)}{\beta}}}{2\vartheta^{\frac{1}{2\beta}}\Delta(1-(H+2\beta))}.$$

Assuming $\beta \ll 1$ we use Stirling's formula to obtain

$$\tilde{\varepsilon}^{net}(b) = \frac{\sqrt{\pi\beta}n(n-1)(\frac{b}{n})^{-\frac{1-(H+2\beta)}{\beta}}}{(2e\beta\vartheta)^{\frac{1}{2\beta}}\Delta(1-(H+2\beta))}. \tag{22}$$

We use the same steps as in Sect. III-C to minimize $\varepsilon^{net}(b)$ over β. For $\beta \ll (1-H)/2$ we find the near optimal solution

$$\beta^* = \frac{n^{2-2H}}{2(-\log \varepsilon_a)}$$

where ε_a is defined in Cor. 1. The term ε_a is used here only to shorten notation and to express ε^{net} as a multiple of ε_a in the

sequel. We emphasize that we do not use the approximation by the largest term (3).

We define $\psi = (1-H)/(1-(H+2\beta))$ and use χ from (13) to find by insertion of β^* into (22) that

$$\tilde{\varepsilon}^{net}(b) = \frac{n-1}{n^H}\left(\frac{H}{1-H}\right)^H \frac{\sqrt{\pi}\psi\sigma b^{1+H}}{\Delta(r^{cr}-\lambda)^{1+H}}\varepsilon_a^{(1+\log\chi)n^{2H-2}}.$$

Note that both χ and ψ approach 1 as $\beta \to 0$. Assuming $\beta^* \ll (1-H)/2$ we can generally find constants that bound χ and ψ from below, respectively, from above. Inserting ε_a as defined in Cor. 1 and using positive constants c_i we obtain

$$\tilde{\varepsilon}^{net}(b) \leq n^{1-H}c_1 b^{1+H}e^{-c_2 b^{2-2H}n^{2H-2}}.$$

We let $b = n(c_0 \log n)^{\frac{1}{2-2H}}$ for $n \geq 2$ and find

$$\tilde{\varepsilon}^{net} \leq (c_0 \log n)^{\frac{1+H}{2-2H}}c_1 n^{2-c_0 c_2}.$$

Generally, there exists $c_0 > 2/c_2$ such that $\tilde{\varepsilon}^{net}$ is upper bounded by a constant for all n. Fixing $\tilde{\varepsilon}^{net}$ it follows that

$$b \in \mathcal{O}\left(n(\log n)^{\frac{1}{2-2H}}\right).$$

Finally, we verify that β^* decreases with n where the decay is proportional to $1/\log(n)$. This confirms the assumption that given β^* is small it also remains small with increasing n. Recall that the above approximation using Stirling's formula is exact for $\beta \to 0$.

For comparison we use the same network, however, with EBB instead of fBm cross traffic. As before, the EBB traffic is an aggregate of m on-off sources. From the EBB sample path envelope in Sect. IV it follows that $S^i(t) = (C - r^{cr})t$ is a leftover service curve with deficit profile $\varepsilon(b) = e^{-\theta b}/(\theta(r^{cr} - m\rho(\theta)))$. After relaxation of the service curve by rate $\delta > 0$ the sample path deficit profile follows from (21) as $\varepsilon_\delta(b) = e^{-\theta b}/(\delta\theta^2(r^{cr} - m\rho(\theta)))$ for any θ that satisfies $m\rho(\theta) < r^{cr}$. The network service curve under EBB cross traffic becomes $S^{net}(t) = (C - r^{cr} - \Delta)t$ with deficit profile

$$\varepsilon^{net}(b) = \frac{1}{r^{cr} - m\rho(\theta)} \inf_x \left\{ \frac{(n-1)^2 e^{-\theta\frac{b-x}{n-1}}}{\Delta\theta^2} + \frac{e^{-\theta x}}{\theta} \right\}$$

where $x \in (0, b)$ and θ such that $m\rho(\theta) < r^{cr}$.

Solving $\varepsilon^{net}(b)$ for b yields that end-to-end performance bounds under EBB cross traffic are in $\mathcal{O}(n \log n)$. This result is derived in [19] and also proven as a lower bound in [36]. In contrast, our result for fBm cross traffic is a scaling in $\mathcal{O}\left(n(\log n)^{\frac{1}{2-2H}}\right)$. The scaling is largely determined by the decay rate of overflow probabilities. The difference to EBB is caused by the Weibullian decay for fBm traffic as opposed to the exponential decay for EBB traffic, see Sect. III-C. We find that the poly-logarithmic scaling component that increases with H is due to LRD. For the special case $H = \frac{1}{2}$ the overflow probability of fBm traffic decays exponentially fast and we recover the scaling $\mathcal{O}(n \log n)$ for EBB cross traffic. While we conclude that LRD has significant impact on single system performance bounds, we find that the additional effect due to concatenation of tandem systems is moderate.
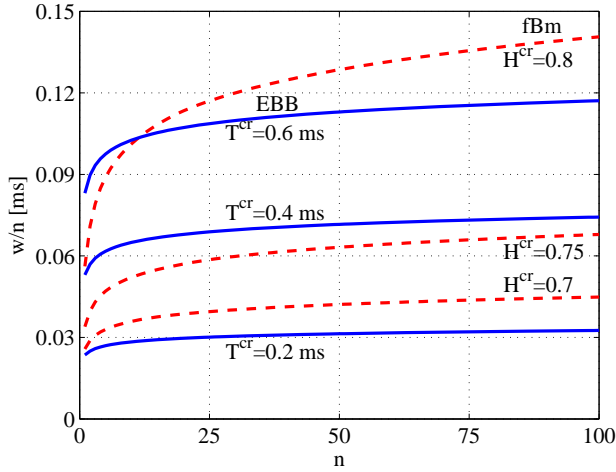
Fig. 9. End-to-end delay bounds as in Fig. 8, however, normalized by n for EBB and fBm cross traffic, respectively. Under EBB cross traffic w/n grows logarithmically with n whereas it grows poly-logarithmically under fBm cross traffic where the exponent grows with H as $1/(2 - 2H)$.

In Fig. 9 we show end-to-end delay bounds for CBR through traffic under either EBB or fBm cross traffic. We use the same parameters and the same network as for Fig. 8, however, we normalize the delay by the number of systems in series n to compare the logarithmic and poly-logarithmic scaling under EBB and fBm cross traffic, respectively. The traffic parameters are given in Tab. II. As before, we optimize the parameters of the traffic envelopes numerically. Fig. 9 clearly confirms the logarithmic growth of the normalized delay bounds. Moreover, Fig. 9 shows that a variation of the parameter T^{cr}, that determines the correlation of the EBB cross traffic, mainly reproduces staggered versions of w/n. In contrast, varying parameter H^{cr} of fBm cross traffic changes the exponent of the poly-logarithmic scaling and hence alters the shape of w/n.

While the scaling result for EBB is in $\Theta(n \log n)$, i.e. it is proven both as an upper bound [19] and a lower bound [36], a lower bound that complements the scaling for fBm in $\mathcal{O}\big(n(\log n)^{\frac{1}{2-2H}}\big)$ still remains to be derived. Note that the scaling results under fBm, respectively, EBB cross traffic are derived without making assumptions about statistical independence of the service left over at individual systems. Under the additional assumption of statistical independence end-to-end performance bounds that scale in $\mathcal{O}(n)$ are derived in [20] for $(\sigma(\theta), \rho(\theta))$ constrained cross traffic [15] that is closely related to the EBB traffic model. The effect of statistical independence on the scaling under fBm cross traffic is not addressed in this work and is an unresolved research question. For an elaboration on known scaling results see [35].

VI. Conclusions

The contribution of this paper are end-to-end statistical performance bounds for a through flow in a network under fBm cross traffic with LRD. To this end, we used the framework of the stochastic network calculus. We developed a sample path envelope for fBm traffic that complements a

known approximation using the largest term. Our sample path envelope and the approximation agree in the Weibullian decay of overflow probabilities. We recovered the previous result at the point in time where the violation of an affine upper envelope by fBm traffic is most probable.

From the derived envelopes we obtained the service curve left over by fBm cross traffic at a system. By convolution of these leftover service curves we derived a network service curve. An essential intermediary result for the successful derivation is the twofold integrability of the overflow probability. We showed a numerical evaluation of the impact of fBm cross traffic on the end-to-end performance of through flows. We proved that end-to-end performance bounds for n systems in series grow in $\mathcal{O}\big(n(\log n)^{\frac{1}{2-2H}}\big)$.

References

[1] L. Kleinrock, *Queueing Systems, Volume 1: Theory*. John Wiley & Sons, 1975.

[2] G. Appenzeller, I. Keslassy, and N. McKeown, "Sizing router buffers," in *Proc. of ACM SIGCOMM*, Sep. 2004, pp. 281–292.

[3] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Netw.*, vol. 2, no. 1, pp. 1–15, Feb. 1994.

[4] V. Paxson and S. Floyd, "Wide-area traffic: The failure of Poisson modeling," *IEEE/ACM Trans. Netw.*, vol. 3, no. 3, pp. 226–244, Jun. 1995.

[5] M. E. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: evidence and possible causes," *IEEE/ACM Trans. Netw.*, vol. 5, no. 6, pp. 835–846, Dec. 1997.

[6] W. Willinger, M. Taqqu, R. Sherman, and D. Wilson, "Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level," *IEEE/ACM Trans. Netw.*, vol. 5, no. 1, pp. 71–86, Feb. 1997.

[7] A. Feldmann, A. C. Gilbert, P. Huang, and W. Willinger, "Dynamics of IP traffic: A study of the role of variability and the impact of control," in *Proc. of ACM SIGCOMM*, Aug. 1999, pp. 301–313.

[8] M. Taqqu, W. Willinger, and R. Sherman, "Proof of a fundamental result in self-similar traffic modeling," *SIGCOMM Comput. Commun. Rev.*, vol. 27, no. 2, pp. 5–23, Apr. 1997.

[9] B. Mandelbrot and J. V. Ness, "Fractional Brownian motions, fractional noises and applications," *SIAM Review*, vol. 10, no. 4, pp. 422–437, 1968.

[10] I. Norros, "A storage model with self-similar input," *Queueing Systems*, vol. 16, no. 3, pp. 387–396, Sep. 1994.

[11] ——, "On the use of fractional Brownian motion in the theory of connectionless networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 953–962, Aug. 1995.

[12] N. Duffield and N. Connell, "Large deviations and overflow probabilities for the general single-server queue, with applications," 1995.

[13] L. Massouli and A. Simonian, "Large buffer asymptotics for the queue with FBM input," *Journal of Applied Probability*, vol. 36, pp. 894–906, 1999.

[14] F. P. Kelly, "Notes on effective bandwidths," ser. Royal Statistical Society Lecture Notes. Oxford University, 1996, no. 4, pp. 141–168.

[15] C. Chang, *Performance Guarantees in Communication Networks*. Springer, 2000.

[16] J.-Y. Le Boudec and P. Thiran, *Network Calculus A Theory of Deterministic Queuing Systems for the Internet*. Springer, 2001.

[17] C. Li, A. Burchard, and J. Liebeherr, "A network calculus with effective bandwidth," *IEEE/ACM Trans. Netw.*, vol. 15, no. 6, pp. 1442–1453, Dec. 2007.

[18] A. Burchard, J. Liebeherr, and S. Patek, "A min-plus calculus for end-to-end statistical service guarantees," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 4105–4114, Sep. 2006.

[19] F. Ciucu, A. Burchard, and J. Liebeherr, "Scaling properties of statistical end-to-end bounds in the network calculus," *IEEE/ACM Trans. Netw.*, vol. 14, no. SI, pp. 2300–2312, Jun. 2006.

[20] M. Fidler, "An end-to-end probabilistic network calculus with Moment Generating Functions," in *Proc. of IWQoS*, Jun. 2006, pp. 261–270.

[21] Y. Jiang and Y. Liu, *Stochastic Network Calculus*. Springer, Sep. 2008.

[22] A. K. Parekh and R. G. Gallager, "A Generalized Processor Sharing approach to flow control in Integrated Services Networks: The single-node case," *IEEE/ACM Trans. Netw.*, vol. 1, no. 3, pp. 344–357, 1993.

[23] H. Sariowan, R. L. Cruz, and G. C. Polyzos, "Scheduling for quality of service guarantees via service curves," in *Proc. of IEEE ICCCN*, Sep. 1995, pp. 512–520.

[24] R. L. Cruz, "Quality of service management in Integrated Services networks," in *Proc. of the Semi-Annual Research Review, Center of Wireless Communication, UCSD*, Jun. 1996.

[25] ——, "A calculus for network delay, part I and II: Network elements in isolation and network analysis," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 114–141, Jan. 1991.

[26] O. Yaron and M. Sidi, "Performance and stability of communication networks via robust exponential bounds," *IEEE/ACM Trans. Netw.*, vol. 1, no. 3, pp. 372–385, Jun. 1993.

[27] D. Starobinski and M. Sidi, "Stochastically Bounded Burstiness for communication networks," *IEEE Trans. Inf. Theory*, vol. 46, no. 1, pp. 206–212, Jan. 2000.

[28] R.-R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn, "Statistical service assurances for traffic scheduling algorithms," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 12, pp. 2651–2664, Dec. 2000.

[29] Q. Yin, Y. Jiang, S. Jiang, and P. Y. Kong, "Analysis of generalized stochastically bounded bursty traffic for communication networks," in *Proc. of IEEE LCN*, Nov. 2002, pp. 141–149.

[30] A. Kumar, D. Manjunath, and J. Kuri, *Communication Networking: An Analytical Approach*. Morgan Kaufmann Publishers Inc., 2004.

[31] P. W. Glynn and W. Whitt, "Logarithmic asymptotics for steady-state tail probabilities in a single-server queue," pp. 131–156, Jan. 1994.

[32] C. Grimm and G. Schlüchtermann, *IP-Traffic Theory and Performance*. Springer, 2008.

[33] G. Mayor and J. Silvester, "Time scale analysis of an ATM queueing system with long-range dependent traffic," in *Proc. of IEEE INFOCOM*, Apr. 1997, pp. 205–212.

[34] N. Fonseca, G. Mayor, and C. Neto, "On the equivalent bandwidth of self-similar sources," *ACM Trans. Model. Comput. Simul.*, vol. 10, no. 2, pp. 104–124, Apr. 2000.

[35] F. Ciucu, "Scaling properties in the stochastic network calculus," Ph.D. dissertation, Univ. of Virginia, Aug. 2007.

[36] A. Burchard, J. Liebeherr, and F. Ciucu, "On Theta (H log H) scaling of network delays," in *Proc. of IEEE INFOCOM*, May 2007, pp. 1866–1874.

APPENDIX

Lemma 1 (Gamma Function). *For $x \in (0,1)$ and $\xi > 0$ it holds that*

$$\int_0^\infty x^{t^\xi} dt = \frac{\Gamma\left(\frac{1}{\xi}\right)}{\xi(-\log x)^{\frac{1}{\xi}}}$$

Proof: The definition of the Gamma function by Euler states that for all $y > 0$

$$\Gamma(y) = \int_0^1 (-\log t)^{y-1} dt.$$

We substitute $t = e^{-z\tau^\xi}$ where $z > 0$ and $\xi > 0$. It follows that $dt = -z\xi e^{-z\tau^\xi}\tau^{\xi-1} d\tau$ and

$$\Gamma(y) = z^y \xi \int_0^\infty \tau^{\xi y - 1} e^{-z\tau^\xi} d\tau.$$

Assuming $\xi > 0$ and letting $y = 1/\xi$ yields

$$\Gamma\left(\frac{1}{\xi}\right) = z^{\frac{1}{\xi}} \xi \int_0^\infty e^{-z\tau^\xi} d\tau.$$

Finally, we substitute $z = -\log x$ where $x \in (0,1)$. ∎